Introduction to Statistical Methods in Political Science

Lecture 12: Small-Sample Inference for Means: Student-t Toolbox

Ignacio Urbina

Ph.D. Candidate

Why Do We Need New Tools?

Motivating example: Tiny Exit Poll

A survey team intercepts **12** early voters leaving a rural precinct. They record time-in-booth (minutes) to study wait-time equity. Longer time-in-booth may signal inefficiencies, understaffing, or barriers to quick voting (e.g., confusing ballots, slow machines).

- Population SD σ is *unknown*.
- Sample histogram shows minor right-skew and an outlier at 14 min.
- Question: Can we still make a reasonably justified inference about the true mean wait time?

Z procedures from last week assume:

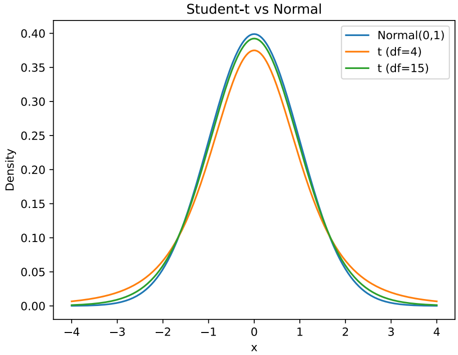
$$\text{Either } n \geq 30 \quad \textbf{or} \quad \sigma \text{ is known.}$$

Neither is true here $\Longrightarrow$ enter Student-t.

What actually changes when n is small?

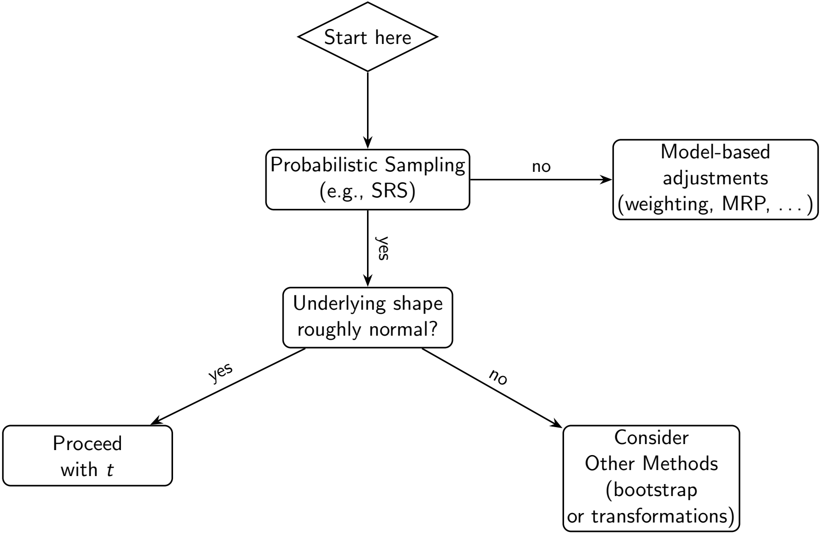
- We swap σ for the noisier estimate s.
- That extra "plug-in" noise fattens the tails of our test statistic.
- We can't rely on the "plug-in principle" (Law of Large Numbers doesn't hold).
- Student-t distribution captures this inflation with **degrees of freedom**:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$



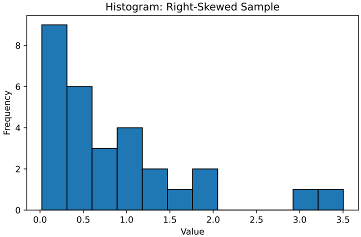
(visual: t_4, t_{15}, and $N(0,1)$)

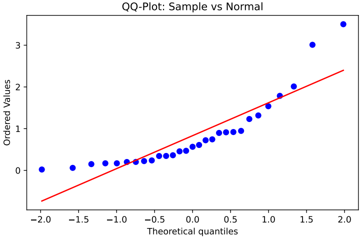
Checklist before using a small-sample t method



Data-shape diagnostics come first

- With $n < 30$ a single high outlier can wreck validity.
- Always inspect a **histogram** and **QQ-plot**.
- Rule of thumb: mild skew is tolerable for $n \geq 15$; heavy skew/outliers call for non-parametrics or resampling.



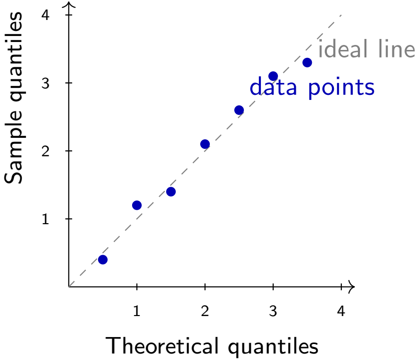


Intuition Behind a Q–Q Plot

What is a Q–Q plot?

- Compares your data's quantiles (y-axis) to theoretical quantiles (x-axis).
- If data follow the chosen distribution, points lie roughly on the 45° line.
- Deviations highlight skew, heavy tails, or outliers.
- Think of "lining up" your sample against the ideal.

QQ-Plot



One-Sample CI for a Mean

Sampling Distribution of Standardized $\bar{x}$ (Small Sample)

Goal: Understand the behavior of $\bar{x}$ as an estimator of μ when sample size is small ($n < 30$).

Common assumptions:

- Data are collected via SRS.
- The population distribution is approximately Normal (key for small n inference).
- Population SD (σ) is unknown.

Result:

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

- Use s to estimate $\sigma \rightarrow$ introduces extra variability.
- The t distribution accounts for this with heavier tails than Normal.

Confidence Interval for μ in Small Samples – The recipe

$$\boxed{\bar{x} \ \pm \ t^*_{df=n-1,\,1-\alpha/2} \times \frac{s}{\sqrt{n}}}$$

- **Degrees of freedom** $df = n - 1$.
- **Critical value** t^* comes from a table or software. It depends on both df and α.
- **Interpretation** follows the familiar "We are 95% confident . . .".

Question – How does t^* compare with z^*?

Suppose $\alpha = 0.05$.

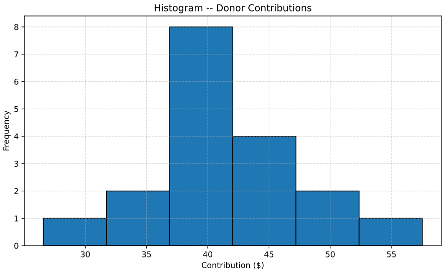
A. t^*_{19} is **smaller** than $z^* = 1.96$
B. t^*_{19} is **equal to** z^*
C. t^*_{19} is **larger** than z^*
D. It cannot be determined because t^* depends on the sample's standard deviation, whereas z^* does not

(Answer: C; heavier tails)

Worked example: Local campaign donors

Goal: assess mean contributions, based on our sample of 18 local donors, by constructing a 90% confidence interval for the mean donation.

- $n = 18$, $\bar{x} = \$42.8$, $s = \$9.2$.
- 90% confidence wanted ($\alpha = 0.10$).



Solution:

$$t^*_{n-1,\,1-\frac{\alpha}{2}} = t^*_{17,\,0.95} = 1.740$$

$$SE = \frac{9.2}{\sqrt{18}} = 2.17$$

$$ME = 1.740 \times 2.17 = 3.77$$

CI:
$\$42.8 \pm 3.8 = (\$39.0,\ \$46.6)$.

Hypothesis Testing for Means

Five–Step Road Map

1. **State H_0 and H_a** (parameter language, direction – one or two tailed)
2. **Choose** α (tolerable Type I risk)
3. **Compute test statistic** $T = \dfrac{\text{estimate} - \text{null}}{SE}$

   $T \sim t_{df}$ if checklist passes.
4. **Decision rule** — compare either $|T|$ to a critical value t^*_{df} *or* use a p-value.
5. **Conclusion in context**. (plain-English, mention evidence strength)

Worked Example — Average Time in Booth

Rural exit-poll revisit.

In response to concerns about unequal voting experiences, electoral officials assert that average time spent in the voting booth should not exceed **6 minutes**. They claim that rural polling stations are operating efficiently and equitably.

Activists, skeptical of this claim, conduct an informal audit by collecting a small **simple random sample** of $n = 12$ early voters at a rural precinct. Each voter is asked how long they spent in the booth, from entry to casting their ballot.

- Sample mean: $\bar{x} = 7.8$ minutes
- Sample standard deviation: $s = 3.1$ minutes
- Officials' claim: $\mu = 6$ minutes (true average time)

Goal: Test whether the true mean booth time μ differs from 6 minutes. Use a two-sided t test at significance level $\alpha = 0.05$.

Worked Example – 5-Step Procedure

Step 1: State hypotheses

$H_0 : \mu = 6$ minutes (official claim)
$H_a : \mu \neq 6$ minutes (activists suspect difference)

Step 2: Set significance level

$\alpha = 0.05$ (two-tailed test). Hence, Critical value: $t^*_{0.975,\,11} = 2.201$.

Step 3: Compute test statistic

Sample mean: $\bar{x} = 7.8$, sample SD: $s = 3.1$, $n = 12$
Standard error: $SE = \frac{3.1}{\sqrt{12}} = 0.90$
Test statistic: $T = \frac{7.8-6}{0.90} = 2.00$

Step 4: Make decision

Degrees of freedom: $df = 12 - 1 = 11$. Critical value: $t^*_{0.975,\,11} = 2.201$
Since $|T| = 2.00 < 2.201$, we **fail to reject** H_0

Step 5: Conclusion in context

Evidence is insufficient (at the 5% level) to conclude that the true average booth time differs from the official 6-minute claim.

Inference for Means in Paired-Samples

Why Use Paired Measurements?

Context: In many research settings, it's hard to detect a treatment effect when individual baseline differences are large.

Solution: Pairing allows each subject (or unit) to serve as their own control.

Common examples of pairing:

- **Before vs. after** a treatment or policy change (e.g., turnout before/after voter ID law)
- **Twin studies** in medical or behavioral research (genetically matched units)
- **Matched groups or regions** — e.g., similar counties, classrooms, or districts

Why Use Paired Measurements?

Why it works:

- Controls for *individual-level variability* (age, baseline attitudes, income, etc.)
- Focuses analysis on the **within-pair difference** d_i
- Turns the problem into a simpler one-sample inference on $\mu_d =$ mean change
- Usually improves precision and statistical power

Sampling Distribution of the Mean Difference

Data: For each of n units we observe a before/after (or matched) pair and compute the difference $d_i = x_{\text{after},i} - x_{\text{before},i}$.

Assumptions:

- Differences d_i are independent draws.
- Distribution of d_i is approximately Normal (key when $n < 30$).

Estimator	**Sampling Distribution**	**Sample SD**
$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$	$\frac{\bar{d} - d}{s_d/\sqrt{n}} \sim t_{df=n-1}$	$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$

- s_d estimates the unknown σ_d, inflating tail thickness.
- Degrees of freedom $df = n - 1$ adjust for that extra noise.

Sampling Distribution of the Mean Difference

Data: For each of n units we observe a before/after (or matched) pair and compute the difference $d_i = x_{\text{after},i} - x_{\text{before},i}$.

Assumptions:

- Differences d_i are independent draws.
- Distribution of d_i is approximately Normal (key when $n < 30$).

Estimator	Sampling Distribution	Sample SD
$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$	$\frac{\bar{d}-d}{s_d/\sqrt{n}} \sim t_{df=n-1}$	$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$

- s_d estimates the unknown σ_d, inflating tail thickness.
- Degrees of freedom $df = n - 1$ adjust for that extra noise.

CI & Test for Mean Difference μ_d

$$\text{Confidence Interval:} \quad \bar{d} \;\pm\; t^*_{df=n-1,\,1-\alpha/2}\,\frac{s_d}{\sqrt{n}}$$

$$\text{Test statistic:} \quad T = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}} \;\sim\; t_{\,df=n-1}$$

- $\mu_{d,0}$ is the *null hypothesized* mean difference (often 0 for "no change").

Key assumptions

- Differences d_i are *independent*.
- Sample size for *df* is the count of *pairs*, not raw observations.
- The distribution of d_i is *approximately Normal* (check histogram/QQ-plot).
- For CI: same as test, plus choice of confidence level $1-\alpha$.

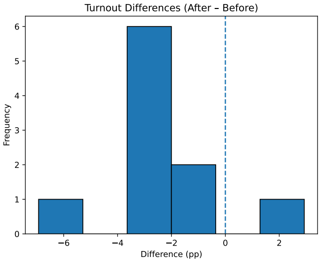
Example: Turnout before vs after voter-ID law

Ten matched counties $\rightarrow$ $n = 10$. $\bar{d} = -1.8$ pp, $s_d = 2.7$ pp; **Task:** Compute 95 % confidence interval.

$$t^*_{9,\,0.975} = 2.262,$$
$$SE = \frac{2.7}{\sqrt{10}} \approx 0.85,$$
$$ME = 2.262 \times 0.85 \approx 1.9$$

CI: $-1.8 \pm 1.9 = (-3.7,\ 0.1)$ pp

Take-away: 95% CI = (-3.7 pp, +0.1 pp): a zero increase can't be excluded.



Example – Interpreting the 95% CI

- CI for mean change: $(-3.7$ pp, 0.1 pp$)$ (pp.=percentage points).
- All values in this range are equally compatible with the data at the 5% significance level
- **You cannot** assign greater "plausibility" to negative vs. positive values
- Conclusion: Data support a decrease, no change, or a small increase — nothing beyond this interval is consistent at 95%

Practice Problem 2 – Paired Data, Small Sample

$$T = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}} \sim t_{n-1}$$

Context

- $n = 9$ individuals measured **before** and **after** a training program.
- Goal: Test if mean improvement μ_d differs from 0.
- Use a **paired** t test — small sample, assume differences $\approx$ normal.

Quick practice (think–pair–share):
$\bar{d} = 4.1,\ s_d = 5.4,\ \alpha = 0.10$ (two-sided).
Reject H_0? $\rightarrow T = 2.27 > t_8^* = 1.86 \rightarrow$ **Yes**.

Addendum – Inference for Paired Differences (Large n)

Scenario: You observe two measurements on each unit (e.g., before vs after treatment) and compute the difference $d_i = x_{\text{after},i} - x_{\text{before},i}$.

When n is large, we invoke the Central Limit Theorem:

$$\bar{d} \sim N\left(\mu_d,\ \frac{\sigma_d^2}{n}\right) \quad \text{(approximately, by CLT).}$$

Use Z procedures if:

- $n \geq 30$ (number of *pairs*),
- Independence: pairs are randomly sampled or randomly assigned,
- You estimate σ_d with sample SD (s_d).

$$\textbf{CI for } \mu_d : \bar{d} \pm z^*_{1-\alpha/2} \cdot \frac{s_d}{\sqrt{n}} \qquad \text{Test stat: } Z = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}}$$

Comparing Two Small Samples

Comparing Means Using Small Samples

Core Question: Do two distinct groups differ *on average* in some key outcome?

- Do 12 rural precincts using new machines have shorter average wait times than 12 using the old ones?
- Do honors students in a pilot class ($n = 14$) outperform regular students ($n = 15$) on a civics quiz?
- Are turnout rates different across two small counties in a special election ($n = 10$ precincts each)?

Comparing Means Using Small Samples

What makes this different from one-sample inference?

- Two samples = two sources of variability
- Independence between groups is critical
- Assumptions about spread (equal vs unequal variance) influence the method

Our goal: Infer whether population means μ_1 and μ_2 are different, using small samples.

Why Variance Matters More with Small Samples

In large samples, we relied on the Law of Large Numbers:

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{(plug-in with } s_1, s_2\text{)}$$

But with small samples:

- Estimates s_1 and s_2 are noisy — not reliable stand-ins for σ_1, σ_2
- This extra uncertainty fattens the tails of our test statistic
- We need to adjust using a **t distribution** — with carefully chosen degrees of freedom

When & How to Pool Variances

Equal-variance assumption $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in the populations. A quick screen: variance (or SD) ratio < 2 *and* similar histograms.

Pooled estimate of the common SD

$$s_{pooled}^2 = s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \qquad s_p = \sqrt{s_p^2}.$$

$$SE_{\text{pooled}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \qquad df = n_1 + n_2 - 2.$$

Use pooled t only when:

- Boxplots / histograms show comparable spread;
- Sample sizes are not wildly unequal;
- A formal test (e.g. Levene) does *not* reject equal variances.

Otherwise, default to Welch's unpooled procedure.

Welch **t** statistic (safer default)

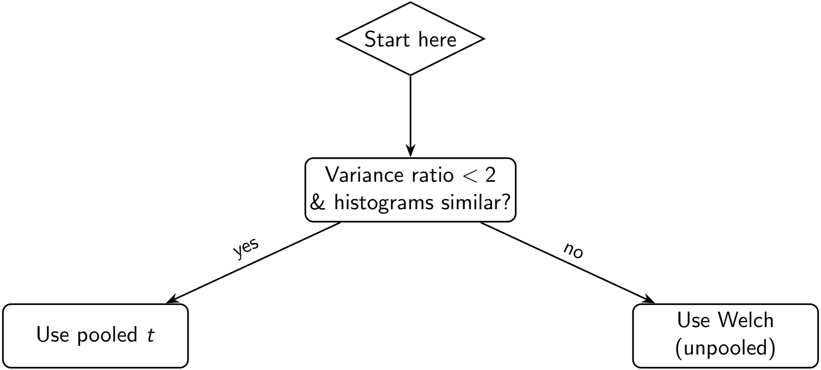
$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \sim \quad t_{df_{\text{Welch}}}$$

Welch degrees of freedom (software reports this)

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Use pooled-SD version only when diagnostics support equal population variances.

Pooled vs. Welch decision tree



Example — Civics Quiz Scores: Honors vs Regular

- **Research context:** Instructor wants to know if an enriched honors curriculum leads to higher civics-quiz performance than the standard curriculum
- **Data:** 20-question multiple-choice quiz (0–100 scale), administered simultaneously to two sections in Spring term
- **Goal:** Estimate and test the difference in true mean scores between honors vs. regular students
- Honors class: $n_1 = 12,\ \bar{x}_1 = 77.3,\ s_1 = 8.4$
- Regular class: $n_2 = 15,\ \bar{x}_2 = 70.1,\ s_2 = 7.1$
- Hypothesis test: $H_0 : \mu_1 - \mu_2 = 0$ vs. two-sided H_a at $\alpha = 0.05$

Calculations

$$SE = \sqrt{\frac{8.4^2}{12} + \frac{7.1^2}{15}} = 3.29, \quad T = \frac{77.3 - 70.1}{3.29} = 2.19$$

$$df_{\text{Welch}} = \frac{(8.4^2/12 + 7.1^2/15)^2}{\frac{8.4^4}{12^2 \cdot 11} + \frac{7.1^4}{15^2 \cdot 14}} \approx 20.7$$

Two-tailed critical value: $t^*_{0.975,\,20} = 2.086$. Since $2.19 > 2.086$ we **reject** H_0.

Conclusion: Honors students score significantly higher ($\approx$ 7 pts).

Common traps with two-sample t

- **Heteroskedasticity**: ignoring unequal variances shrinks SE.
- **Imbalanced** n: smaller group sample size drives df; watch power.
- **Multiple testing**: comparing many sub-groups inflates Type I error (adjust α or use FDR).

Key Formulas and Takeaways

Cheat-Sheet – Small Sample Inference

Scenario	**Confidence Interval**	**Test Statistic** (Δ_0 or μ_0 in numerator)
One mean μ	$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$	$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (df = n - 1)$
Paired mean μ_d	$\bar{d} \pm t^*_{n-1} \frac{s_d}{\sqrt{n}}$	$T = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}} \quad (df = n - 1)$
Two means $\mu_1 - \mu_2$ (Welch)	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (df = \text{Welch})$
Two means $\mu_1 - \mu_2$ (Pooled)	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (df = n_1 + n_2 - 2)$

- Always check independence *and* approximate Normal shape.
- Use Welch unless equal-variance assumption is defensible.
- Report *df* and *p*-value to two decimals.